The convex optimization landscape of neural networks and the convex geometry of back propagation

Yifei Wang

Stanford University

May 5th

Joint work with Jonathan Lacotte and Prof. Mert Pilanci

Roadmap

- The hidden convex optimization landscape of regularized two-layer ReLU networks[1]

  All globally optimal ReLU neural networks can be found via convex optimization

- Implicit regularization of gradient flow in training two-layer ReLU networks with no regularization[2]

  Unregularized non-convex gradient flow (i.e., backpropagation) converges to an optimal solution of our convex program

[1] Y. Wang, J. Lacotte, M. Pilanci. The Hidden Convex Optimization Landscape of Two-Layer ReLU Neural Networks: an Exact Characterization of the Optimal Solutions. International Conference on Learning Representations (ICLR), 2022 (oral presentation).

[2] Y. Wang, M. Pilanci. The Convex Geometry of Backpropagation: Neural Network Gradient Flows Converge to Extreme Points of the Dual Convex Program. International Conference on Learning Representations (ICLR), 2022 (poster presentation).

Neural networks

- Neural networks exhibit extraordinary optimization and generalization abilities.
- The nonconvex training problem and nonlinear structure of neural networks make our understanding difficult.

Regularized training problem

- Data: $\mathbf{X} \in \mathbb{R}^{n \times d}$ label: $\mathbf{y} \in \mathbb{R}^n$
- Consider the regularized training problem:

$$p_{\text{noncvx}} := \min_{\mathbf{W}_1, \mathbf{w}_2} \left\{ \ell\left(\sum_{i=1}^m (X\mathbf{w}_{1,i})_+ w_{2,i}, \mathbf{y} \right) + \frac{\beta}{2}(\|\mathbf{W}_1\|_F^2 + \|\mathbf{w}_2\|_2^2) \right\}.$$

- $\ell(\mathbf{z}, \mathbf{y})$ is assumed to be a convex function of $\mathbf{z}$. (e.g., logistic, hinge, squared loss)

Convex optimization formulation

- In recent work[1], an optimal neural network can be constructed based on a solution of the convex program

$$
\begin{aligned}
p_{\text{convex}} := \min_{(\mathbf{u}_i, \mathbf{u}_i')_{i=1}^p} \quad & \left\{ \ell\Big(\sum_{i=1}^p \mathbf{D}_i \mathbf{X}(\mathbf{u}_i - \mathbf{u}_i'), \mathbf{y} \Big) + \beta \sum_{i=1}^p (\|\mathbf{u}_i\|_2 + \|\mathbf{u}_i'\|_2) \right\}, \\
\text{s.t.} \quad & (2\mathbf{D}_i - \mathbf{I}_n)\mathbf{X}\mathbf{u}_i \geq 0, (2\mathbf{D}_i - \mathbf{I}_n)\mathbf{X}\mathbf{u}_i' \geq 0, i \in [p].
\end{aligned}
$$

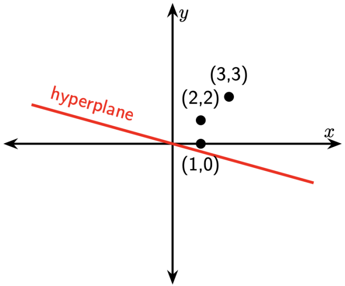
where $\mathbf{D}_1, \dots, \mathbf{D}_p$ are the enumeration of all possible hyperplane arrangements

$$\{\text{diag}(\mathbf{1}(\mathbf{X}\mathbf{u} \geqslant 0)) | \mathbf{u} \in \mathbb{R}^d\}.$$

[1] Mert Pilanci and Tolga Ergen. Neural networks are convex regularizers: Exact polynomial-time convex optimization formulations for two-layer networks. ICML2020.

Hyperplane arrangements

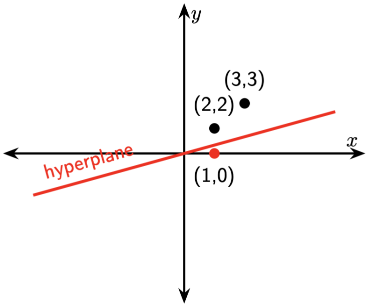
- $n = 3$ samples in $\mathbb{R}^d$, $d = 2$. $X = \begin{bmatrix} x_1^T \\ x_2^T \\ x_3^T \end{bmatrix} = \begin{bmatrix} 3 & 3 \\ 2 & 2 \\ 1 & 0 \end{bmatrix}, y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}.$



$$D_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, D_1 X = \begin{bmatrix} 3 & 3 \\ 2 & 2 \\ 1 & 0 \end{bmatrix}.$$

Hyperplane arrangements

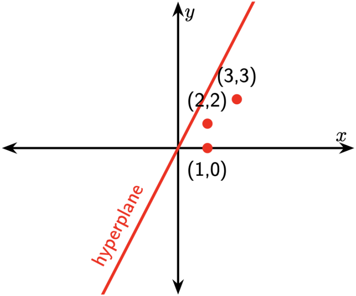
- $n = 3$ samples in $\mathbb{R}^d$, $d = 2$. $X = \begin{bmatrix} x_1^T \\ x_2^T \\ x_3^T \end{bmatrix} = \begin{bmatrix} 3 & 3 \\ 2 & 2 \\ 1 & 0 \end{bmatrix}, y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$.



$$D_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, D_2 X = \begin{bmatrix} 3 & 3 \\ 2 & 2 \\ 0 & 0 \end{bmatrix}.$$

Hyperplane arrangements

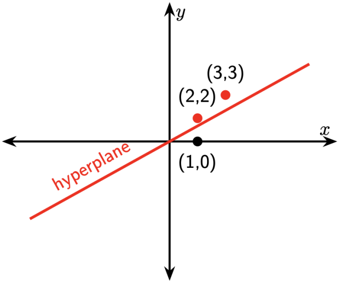
- $n = 3$ samples in $\mathbb{R}^d$, $d = 2$. $X = \begin{bmatrix} x_1^T \\ x_2^T \\ x_3^T \end{bmatrix} = \begin{bmatrix} 3 & 3 \\ 2 & 2 \\ 1 & 0 \end{bmatrix}, y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$.



$$D_3 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, D_3 X = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

Hyperplane arrangements

- $n = 3$ samples in $\mathbb{R}^d$, $d = 2$. $X = \begin{bmatrix} x_1^T \\ x_2^T \\ x_3^T \end{bmatrix} = \begin{bmatrix} 3 & 3 \\ 2 & 2 \\ 1 & 0 \end{bmatrix}, y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$.



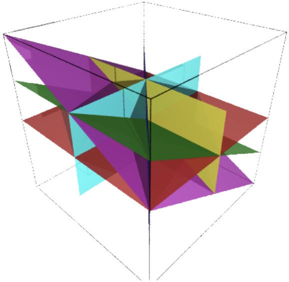
$$D_4 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, D_4 X = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \end{bmatrix}.$$

Upperbound on the number of hyperplane arrangement patterns

- For $\mathbf{X} \in \mathbb{R}^{n\times d}$, $p = \#\{\mathbf{1}(\mathbf{X}\mathbf{u} \geqslant 0) | \mathbf{u} \in \mathbb{R}^d\}$ is bounded by

$$p \leq 2r\left(\frac{e(n-1)}{r}\right)^r,$$

where r is the rank of $\mathbf{X}$.[1]

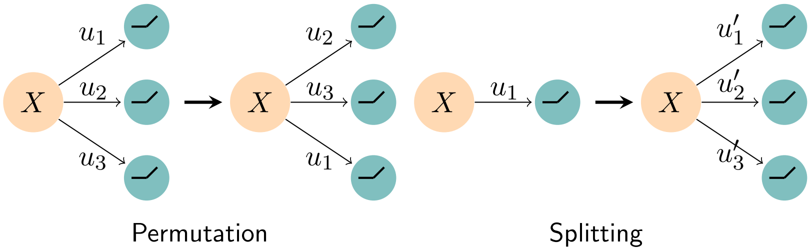


[1] Thomas M Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. IEEE transactions on electronic computers. 1965.

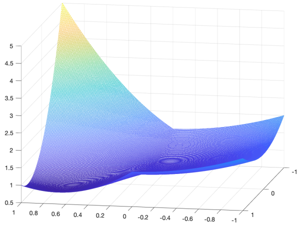
All global optima

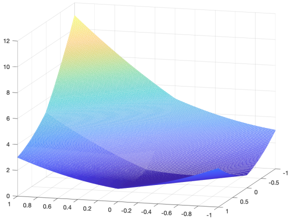
Theorem

Assume that $m \geq m^*$, where $m^* \leq n+1$ is a critical threshold. All optimal solution of p_{noncvx} can be found from the optimal solutions of p_{convex} up to permutation and splitting.



Nonconvex landscape and convex landscape





Comparison of the non-convex landscape (left) and the convex landscape (right). Toy example with data $X = 1$, label $y = 1$ and the ℓ_2 loss. The nonconvex objective is $\mathcal{L}_\beta(u, \alpha) = (1 - \max\{u, 0\}\,\alpha)^2 + \frac{1}{2}(|u|^2 + |\alpha|^2)$. The convex objective is then $\mathcal{L}^c_\beta(v, w) = (1 - v + w)^2 + (|v| + |w|)$ subject to $v, w \geqslant 0$.

Clarke stationary point

- Denote $\mathcal{L}_\beta(\theta)$ as the objective of the nonconvex problem.
- Clarke's subdifferential:

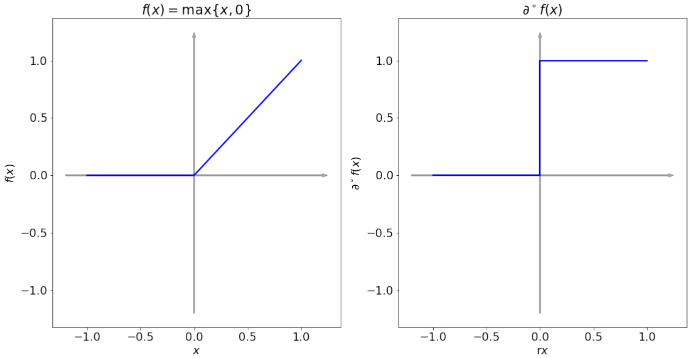
$$\partial^\circ \mathcal{L}_\beta(x) = \mathbf{Co}\left\{\lim_{k\to\infty} \nabla \mathcal{L}_\beta\left(x_k\right) \mid x_k \to x, x_k \in D, \lim_{k\to\infty} \nabla \mathcal{L}_\beta\left(x_k\right) \text{ exists } \right\}$$

- Clarke stationary point:

$$\theta : 0 \in \partial^\circ \mathcal{L}(\theta),$$

- Any local minimizer of $\mathcal{L}_\beta$ is a Clarke stationary point.
- The limit points of SGD are almost surely Clarke stationary with respect to the nonconvex problem.

$$f(x) = \max\{x, 0\},\ \partial^\circ f(x) = \begin{cases} 1, & x > 0, \\ [0,1], & x = 0, \\ 0, & x < 0. \end{cases}$$



Characterization of Clarke stationary point

Theorem

Suppose that $\theta = (\mathbf{W}_1, \mathbf{w}_2)$ is a Clarke's stationary point of the nonconvex problem. Then, θ corresponds to a global optimum of the subsampled convex program:

$$\min_{(\mathbf{u}_i,\mathbf{u}_i')_{i\in\mathcal{I}}} \ell\Big(\sum_{i\in\mathcal{I}} \mathbf{D}_i\mathbf{X}(\mathbf{w}_i - \mathbf{w}_i'), \mathbf{y}\Big) + \beta\sum_{i\in\mathcal{I}}(\|\mathbf{w}_i\|_2 + \|\mathbf{w}_i'\|_2),$$

$$\textit{s.t. } (2\mathbf{D}_i - \mathbf{I}_n)\mathbf{X}\mathbf{w}_i \geq 0, (2\mathbf{D}_i - \mathbf{I}_n)\mathbf{X}\mathbf{w}_i' \geq 0, i \in \mathcal{I},$$

where $\mathcal{I} = \{i \in [p] |$ *there exists* $k \in [m]$ *s.t.* $D_i = \mathrm{diag}(\mathbb{I}(Xu \geq 0))\}$.

Convex optimization formulation and gradient flow

- Simple algorithms including (stochastic) gradient descent minimize the training loss.
- Gradient descent methods serve as heuristics to solve the convex program.
- What kind of solutions will gradient descent/gradient flow find?

Implicit regularization

- For neural network with structures, gradient flow/gradient descent has implicit regularization.
- Classification problem with logistic loss.
- For linear model, the gradient descent maximizes the margin.

$$\arg\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{2}\|\mathbf{w}\|_2^2, \text{ s.t. } y_i\mathbf{w}^T\mathbf{x}_i \geq 1, i \in [n].$$

Problem setting

- Two-layer neural networks with ReLU activation, i.e.,

$$f(\boldsymbol{\theta}, \mathbf{X}) = (\mathbf{X}\mathbf{W}_1)_+\mathbf{w}_2,$$

  where $\mathbf{W}_1 \in \mathbb{R}^{d\times m}$, $\mathbf{w}_2 \in \mathbb{R}^m$ and $\boldsymbol{\theta} = (\mathbf{W}_1, \mathbf{w}_2)$.
- Training problem

$$\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) =: \sum_{i=1}^{n} l(y_i f(\boldsymbol{\theta}; \mathbf{x}_i)),$$

  where $l(q) = \log(1 + \exp(-q))$ is the logistic loss.

Gradient descent and gradient flow

- The gradient descent takes the update rule

$$\boldsymbol{\theta}(t+1) = \boldsymbol{\theta}(t) - \eta(t)\mathbf{g}(t),$$

where $\mathbf{g}(t) \in \partial^\circ \mathcal{L}(\boldsymbol{\theta}(t))$ and ∂° represents the Clarke's subdifferential.

- For gradient flow, the trajectory of the parameter is an arc $\boldsymbol{\theta} : [0, +\infty) \to \Theta = \{(\mathbf{W}_1, \mathbf{w}_2) | \mathbf{W}_1 \in \mathbb{R}^{d\times m}, \mathbf{W}_2 \in \mathbb{R}^m\}$, which satisfies

$$\frac{d}{dt}\boldsymbol{\theta}(t) \in -\partial^\circ \mathcal{L}(\boldsymbol{\theta}(t)),$$

for $t \geq 0$, a.e..

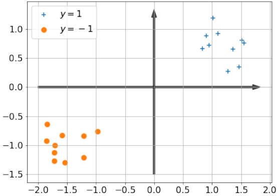
Implicit regularization for homogeneous network

- Assume that there exists time t_0 such that $\mathcal{L}(\theta(t_0)) < 1$, i.e., the data is separated at time t_0.
- Lyu and Li[1] show that with $t \to \infty$, any limiting point of $\frac{\boldsymbol{\theta}(t)}{\|\boldsymbol{\theta}(t)\|_2}$ is along the direction to the KKT point of the max-margin problem

$$\min \frac{1}{2}\|\boldsymbol{\theta}\|_2^2, \text{ s.t. } y_i f(\boldsymbol{\theta}; \mathbf{x}_i) \geq 1, i \in [n].$$

  where $\|\boldsymbol{\theta}\|_2^2 = \|\mathbf{W}_1\|_F^2 + \|\mathbf{w}_2\|_2^2$.
- This is a **nonconvex** optimization problem.
- Does gradient flow converge to a global minimizer?

[1] Lyu, K. and Li, J. (2019). Gradient descent maximizes the margin of homogeneous neural networks.



Theorem

Suppose that $(\mathbf{X}, \mathbf{y}) \in \mathbb{R}^{n\times d} \times \{-1, 1\}^n$ *is orthogonally separable, i.e., for all* $i, i' \in [n]$,

$$\mathbf{x}_i^T \mathbf{x}_{i'} > 0, \text{ if } y_i = y_{i'},$$
$$\mathbf{x}_i^T \mathbf{x}_{i'} \leq 0, \text{ if } y_i \neq y_{i'}.$$

Consider the non-convex subgradient flow applied to the non-convex problem. Suppose that the initialization is sufficiently close to the origin and scaled. Then, the non-convex subgradient flow converges to the global optimum of the non-convex problem up to scaling.

Convex max-margin problem

- The non-convex max-margin problem is equivalent to the following convex program

$$
\begin{aligned}
P^*_{\text{cvx}} = \min \; & \sum_{j=1}^{p} (\|\mathbf{u}_j\|_2 + \|\mathbf{u}'_j\|_2), \\
\text{s.t. } \; & \mathbf{Y} \sum_{j=1}^{p} \mathbf{D}_j \mathbf{X}(\mathbf{u}'_j - \mathbf{u}_j) \geq \mathbf{1}, \\
& (2\mathbf{D}_j - I)\mathbf{X}\mathbf{u}_j \geq 0, (2\mathbf{D}_j - I)\mathbf{X}\mathbf{u}'_j \geq 0, \forall j \in [p].
\end{aligned}
$$

Here $\mathbf{Y} = \text{diag}(\mathbf{y})$.

KKT point

Theorem

The KKT point $(\mathbf{W}_1, \mathbf{w}_2, \boldsymbol{\lambda})$ *of the non-convex max-margin problem corresponds to a KKT point of the convex max-margin problem if and only if* $\boldsymbol{\lambda}$ *satisfies*

$$\max_{\mathbf{u}:\|\mathbf{u}\|_2\leq 1} |\boldsymbol{\lambda}^T(\mathbf{X}\mathbf{u})_+| \leq 1.$$

Equivalently, the variable $\boldsymbol{\lambda}$ *satisfies that for all* $j \in [p]$,

$$\max_{\|\mathbf{u}\|_2\leq 1, (2\mathbf{D}_j - I)\mathbf{X}\mathbf{u}\geq 0} |\boldsymbol{\lambda}^T\mathbf{D}_j\mathbf{X}\mathbf{u}| \leq 1.$$

Dual problem

- The dual problem is given by

$$D^* = \max_{\boldsymbol{\lambda}} \mathbf{y}^T \boldsymbol{\lambda} \text{ s.t. } \mathbf{Y}\boldsymbol{\lambda} \succeq 0, \max_{\mathbf{u}:\|\mathbf{u}\|_2 \leq 1} |\boldsymbol{\lambda}^T (\mathbf{X}^T \mathbf{u})_+| \leq 1.$$

- Suppose that $\boldsymbol{\lambda}^*$ is the optimal dual variable. Then, any optimal primal variable $\mathbf{u}$ belongs to the set

$$\arg\max_{\mathbf{u}:\|\mathbf{u}\|_2 \leq 1} |(\boldsymbol{\lambda}^*)^T (\mathbf{X}^T \mathbf{u})_+|.$$

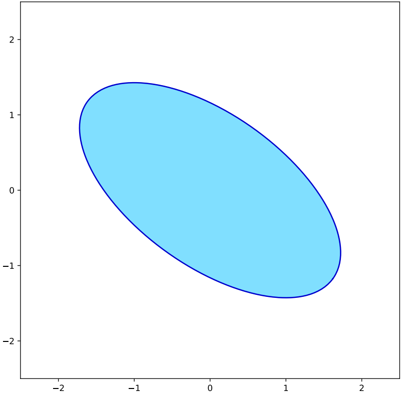
Geometric interpretation

- Geometric interpretation of

$$\max_{\mathbf{u}:\|\mathbf{u}\|_2\leq 1} |\boldsymbol{\lambda}^T(\mathbf{X}^T\mathbf{u})_+|.$$

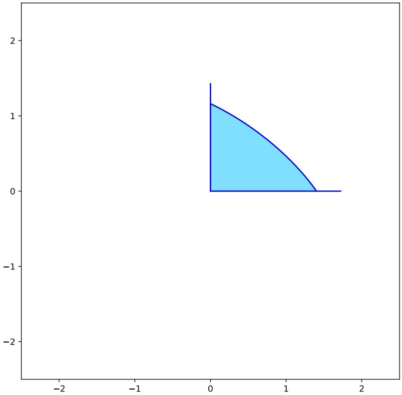
Geometric Interpretation

- Ellipsoid $= \{\mathbf{X}\mathbf{u} : \|\mathbf{u}\|_2 \leq 1\}$.



Geometric Interpretation

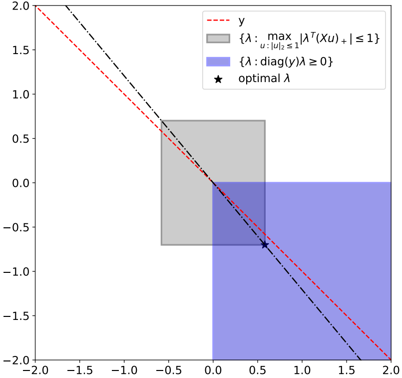
- Rectified Ellipsoid $\mathcal{Q} := \{(\mathbf{X}\mathbf{u})_+ : \|\mathbf{u}\|_2 \leq 1\}$ and its extreme points (spikes).



Geometric Interpretation

- Polar set $\mathcal{Q}^*$ of the Rectified Ellipsoid:

$$\mathcal{Q}^* = \{\boldsymbol{\lambda} : \max_{\mathbf{z}\in\mathcal{Q}} |\boldsymbol{\lambda}^T\mathbf{z}| \le 1\} = \{\boldsymbol{\lambda} : \max_{\mathbf{u}:\|\mathbf{u}\|_2\le 1} |\boldsymbol{\lambda}^T(\mathbf{X}^T\mathbf{u})_+| \le 1\}.$$



Proposition

Suppose that $(\mathbf{X}, \mathbf{y})$ *is orthogonal separable. Suppose that the KKT point* $(\mathbf{W}_1, \mathbf{w}_2, \boldsymbol{\lambda})$ *of the non-convex problem include two neurons* $(\mathbf{w}_{1,i_+}, w_{2,i_+})$ *and* $(\mathbf{w}_{1,i_-}, w_{2,i_-})$ *such that*

$$\mathbb{I}(\mathbf{X}\mathbf{w}_{1,i_+} > 0) \geq \mathbb{I}(y = 1), \quad \mathbb{I}(\mathbf{X}\mathbf{w}_{1,i_-} > 0) \geq \mathbb{I}(y = -1).$$

Then, the dual variable $\boldsymbol{\lambda}$ *satisfies*

$$\max_{\mathbf{u}:\|\mathbf{u}\|_2 \leq 1} |\boldsymbol{\lambda}^T (\mathbf{X}\mathbf{u})_+| \leq 1.$$

In other words, $(\mathbf{W}_1, \mathbf{w}_2)$ *globally minimizes the non-convex max-margin problem.*

Theorem

Consider the training problem for any dataset. Suppose that the neural network is scaled at initialization such that $\|\mathbf{w}_{1,i}\|_2 = |w_{2,i}|$ *for* $i \in [m]$*. Consider the subgradient flow applied to the non-convex problem. Let* $\delta \in (0,1)$*. Suppose that the initialization is sufficiently close to the origin. For random initialization and* $s \in \{-1, 1\}$*, there exist* $T = T(\delta)$ *and neuron* $(\mathbf{w}_{1,i}, w_{2,i})$ *such that*

$$\cos\angle\left(\mathbf{w}_{1,i}(T), s\mathbf{X}^T\mathbf{D}(\mathbf{w}_{1,i}(T))\mathbf{y}\right) \geq 1 - \delta.$$

Here $\mathbf{D}(\mathbf{u}) = \mathrm{diag}(\mathbb{I}(\mathbf{X}\mathbf{u} > 0))$*.*

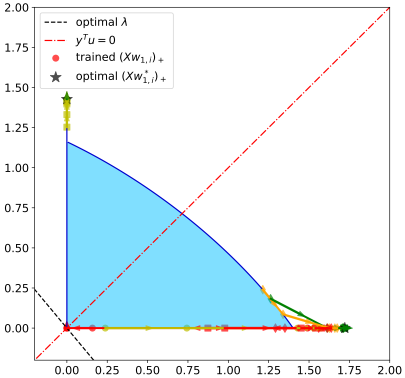


Figure: Trajectories of $(\mathbf{X}\hat{\mathbf{w}}_{1,i})_+$ along the training dynamics of gradient descent.

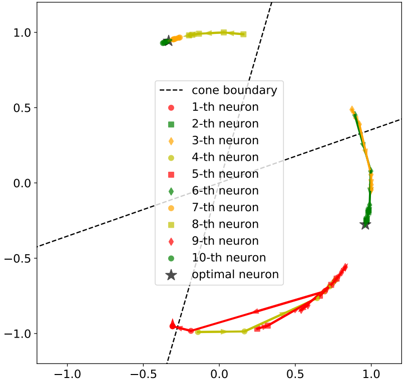


Figure: Trajectories of $\hat{\mathbf{w}}_{1,i} = \frac{\mathbf{w}_{1,i}}{\|\mathbf{w}_{1,i}\|_2}$ along the training dynamics of gradient descent.

Conclusion

- The global optima of the non-convex training problem is given by the optimal set of a cone-constrained convex program.
- Non-convex subgradient flow of the logistic loss can globally maximize the margin of two-layer ReLU networks on orthogonally separable datasets.

Future work

- Characterize the globally optimal set of deep neural networks.
- Study the generalization property of the global optima.
- Extend the analysis to gradient descent training dynamics.
- Extend the analysis to linear separable datasets.